29th Computer Vision Winter Workshop
Jindřichův Hradec, Czech Republic, February 9–12, 2026

Exploring Multimodal Large Language Models for Morphing Attack Detection

Nikola Marić
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
nikola.maric@ijs.si

Marija Ivanovska, Vitomir Štruc
Faculty of Electrical Engineering, University of Ljubljana
Tržaška c. 25, 1000 Ljubljana
{marija.ivanovska, vitomir.struc}@fe.uni-lj.si

Abstract

Existing single-image morphing attack detection (S-MAD) systems often suffer from poor cross-dataset generalization and operate as opaque "black boxes," which is particularly problematic in high-stakes border control scenarios. This paper investigates the adoption of open-source Multimodal Large Language Models (MLLMs) for S-MAD under strict cross-dataset evaluation through two different approaches. First, we assess selected MLLMs in zero-shot settings using a structured forensic prompting framework, which elicits multi-step semantic analysis with human-readable regional attributions. Second, leveraging the lightweight and parameter-efficient LoRA approach and a synthetic training dataset of morphs, we adapt the best-performing MLLM to the morphing attack detection (MAD) task in an efficient, generalizable, and privacy-preserving manner, enhancing the model's sensitivity to diverse morphing artifacts. Our experimental results show that the proposed prompting strategy significantly improves overall attack detection accuracy compared to naive prompting. Moreover, our LoRA-adapted MLLM, Gemma-3 12B, achieves an average equal error rate (EER) of 14.81% *across various morphing attack benchmarks, outperforming widely used MAD models.*

1. Introduction

Face-morphing attacks pose a serious threat to the integrity of biometric security systems by blending facial images of two individuals into a single composite image that simultaneously represents both identities, as illustrated in Figure 1 [15, 19]. By embedding such a morphed image into an identity document, such as a passport, an attacker and an accomplice can jointly and repeatedly bypass automated face matching systems during identity verification [7]. This fundamental vulnerability has motivated the development of dedicated morphing attack detection (MAD) techniques aimed at reliably distinguishing bona fide facial images from morphed ones [7, 9, 14, 21].

MAD methodologies are generally categorized into differential and single-image approaches. Differential MAD methods assess the authenticity of a presented face image by directly comparing it to a trusted reference image acquired during enrollment [6, 33]. In contrast, single-image MAD operates solely on the probe image, making it particularly suitable for practical real-world border control scenarios where reliable reference images are often unavailable, outdated, or of insufficient quality [7]. In this paper, we limit our focus to the more challenging single-image morphing attack detection (S-MAD) task, which poses stricter constraints on available information.

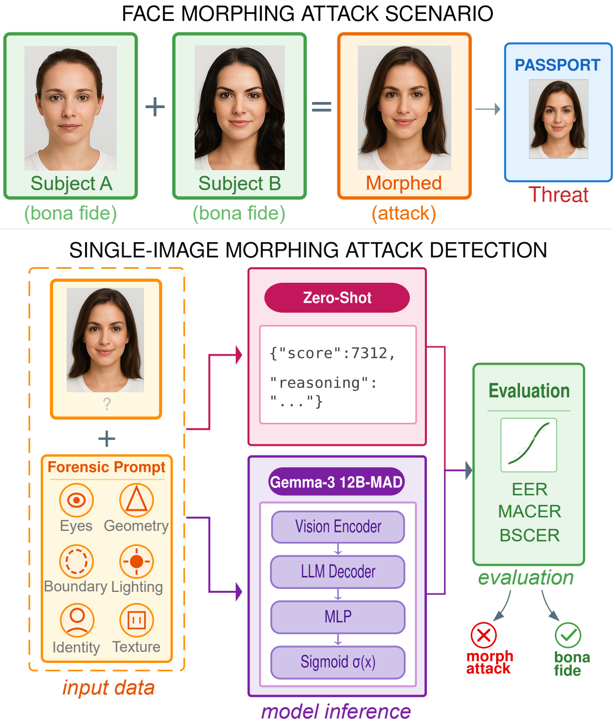


Figure 1. Face morphing attack threat and proposed detection pipeline: Blended facial identities compromise document security (top); our framework performs detection through zero-shot structured forensic analysis or fine-tuned classification (bottom).

While traditional S-MAD detectors rely on handcrafted features, modern approaches leverage deep learning supervised classifiers, which often suffer from severe generalization issues. Recent studies show cross-dataset equal error rates degrading to near-random performance when models are trained on one type of morphing attack techniques and tested on another [11, 14]. This effect is caused not just by the domain shift, but also due to the variability of morphing attack artifacts that characterize individual morphing techniques [22]. This issue has been tackled with the development of different unsupervised [14, 16, 20] and self-supervised MAD techniques [21, 22], but these methods either fail to learn a well-defined boundary between bona fides and morphs, or lack human-interpretable results of the decision [39].

Recent advances in foundation models, spanning vision-only architectures (e.g., ViT), vision–language models (e.g., CLIP), and Multimodal Large Language Models (MLLMs) capable of more complex reasoning, offer a promising route to simultaneously improve MAD generalization and interpretability. General-purpose vision–language models, such as CLIP, have previously shown competitive MAD performance when adapted to the downstream MAD task [7], but they lack explicit semantic reasoning, which is important for forensic analysis in security-critical applications. Conversely, proprietary MLLMs such as GPT-4 have demonstrated both impressive zero-shot detection capabilities and decision interpretability, as reported in preliminary MAD studies [1, 39]. However, their closed-source nature limits its reproducibility, transparency, and task-specific adaptation, motivating the exploration of open-source MLLMs for morphing attack detection. These observations leave two fundamental questions unanswered: *i)* whether open-source MLLMs can match or even exceed specialized CNN-based MAD systems under strict cross-dataset evaluation, and *ii)* whether parameter-efficient adaptation on privacy-friendly synthetic data can equip such models with robust and explainable morphing detection capabilities without overfitting to specific attack processes.

In this paper, we address these questions through a systematic investigation of widely used open-source MLLMs for single-image morphing attack detection, focusing on reproducible, locally deployable architectures rather than proprietary APIs. Our contributions are multifold:

- We conduct the first systematic zero-shot and cross-dataset benchmarking of open-source MLLMs across various, i.e. landmark-, GAN-, and diffusion-based morphing attacks, revealing pronounced differences in their latent forensic sensitivity to MAD.
- We introduce a structured multi-step forensic prompting protocol that leverages chain-of-thought (CoT) reasoning to substantially improve zero-shot morphing attack detection performance over naive prompting, while simultaneously providing interpretable, region-level semantic attributions of detected morphing artifacts.
- We propose a self-supervised, parameter-efficient MLLM adaptation strategy that leverages privacy-preserving synthetic data to achieve strong cross-dataset generalization and competitive performance against widely used state-of-the-art MAD methods.

The remainder of the paper is organized as follows. Section 2 reviews related work on MADs. Section 3 describes our proposed MAD approach. Sections 4 and 5 present experiments and results. Section 6 concludes the paper.

2. Related Work

Face morphing attacks have evolved from early landmark-based warping techniques to increasingly sophisticated generative approaches. Modern attacks span classical landmark-based morphs [13, 28], GAN-based synthesis [8, 38], and more recent diffusion-based morphs [3, 11]. This rapid progression has, in turn, driven the evolution of Morphing Attack Detection (MAD) methodologies, which have advanced from hand-crafted forensic features to supervised deep learning approaches, and more recently to generalized unsupervised frameworks and foundation model-based solutions.

Early Single-Image MAD. Early S-MAD approaches relied on hand-crafted texture descriptors and image forensics. Techniques employing Local Binary Patterns (LBP), Binarized Statistical Image Features (BSIF), and Photo Response Non-Uniformity (PRNU) analysis were effective at identifying blending artifacts or sensor noise inconsistencies [12, 27, 32]. In parallel, Image Quality Assessment (IQA) strategies, such as MagFace [23] and CNNIQA [16], leveraged the observation that morphing processes often degrade facial utility or natural image statistics. These methods established important baselines and remain relevant as quality-based detectors in our comparative evaluation. However, their reliance on low-level cues limits robustness against high-quality, seamless morphs produced by advanced generation techniques, resulting in poor generalization across datasets [31].

Supervised CNN-based MAD. The advent of deep learning shifted the focus toward supervised Convolutional Neural Networks (CNNs) as the dominant paradigm for morphing attack detection. Architectures such as MixFaceNet-MAD [4] and adaptations of Inception and ResNet [17, 35] demonstrated high intra-dataset detection accuracy under controlled laboratory conditions. To further enhance interpretability and spatial precision, Pixel-Wise MAD (PW-MAD) introduced explicit pixel-level supervision to localize morphing regions at fine granularity and provide more transparent decision cues [9]. Despite achieving strong detection performance on known attack types, supervised methods remain prone to severe overfitting to training distributions and dataset-specific artifacts. As a result, they often fail when confronted with previously unseen morphing techniques, such as diffusion-generated morphs, particularly when trained exclusively on landmark-based examples [14]. This lack of robustness highlights a fundamental limitation of CNN-based detectors, whose performance can degrade substantially when exposed to novel and rapidly evolving attack generation mechanisms encountered in real-world deployments.

Unsupervised and Self-Supervised MAD. To address the generalization gap, recent research has pivoted toward the unsupervised and self-supervised learning paradigms, where MADs are trained on bona fide data only, treating morphs as out-of-distribution samples. Self-Paced Learning MAD (SPL-MAD) [14] and MAD-DDPM [20], for example, train reconstruction models that, based on the reconstruction error during testing time flag, morphs as statistical outliers. To improve the estimation of the bona fide distribution, approaches such as OrthoMAD [25] and IDistill [5] optimize their models by simultaneously performing identity disentanglement. Some recent works, e.g., SelfMAD [21] and SelfMAD++ [22], also leverage self-supervised signals, that train the model in a binary manner, by utilizing synthetic morphs that represent typical morphing artifacts, created using augmentations of bona fide data. These methods have significantly reduced cross-dataset error rates by learning generic definitions of face authenticity rather than memorizing specific attack signatures. Our LoRA-based adaptation of MLLMs follows a similar path, as training is performed exclusively on bona fide images with online generation of synthetic artifacts to preserve generalization to unseen attacks.

Foundation Models for MAD. Most recently, the emergence of foundation models has opened a new frontier in MAD. *Caldeira et al.* proposed MADation [7], which fine-tunes the CLIP vision-language model using Low-Rank Adaptation (LoRA) [18], achieving state-of-the-art generalization by leveraging broad pre-trained visual knowledge. Concurrently, *Caldeira et al.* introduced MAD-Prompts [6], exploring multi-prompt aggregation for zero-shot MAD with proprietary MLLMs. However, their study remains limited to closed-source APIs and does not investigate parameter-efficient adaptation or open-weights models. Furthermore, zero-shot evaluations using Multimodal Large Language Models (MLLMs) like GPT-4 Vision have shown that these models possess inherent forensic capabilities, offering both detection and textual explanations without task-specific training [1, 39].

In contrast to MADation, which adapts only the CLIP vision-language model without explicit reasoning mechanisms, our work employs MLLMs that integrate vision and language for semantic analysis. We introduce a structured multi-step Chain-of-Thought forensic prompting protocol for zero-shot MAD. Moreover, we implement self-supervised LoRA fine-tuning of MLLMs applied to both the vision and language components of the models, for their adaptation to the downstream MAD task. Unlike existing MAD methods, our approach provides both interpretability through structured reasoning and cross-dataset performance across diverse morph types.

3. Methodology

In this section, we present two distinct options related to the adoption of open-source MLLMs for MAD. First, we propose a *zero-shot forensic prompting strategy* designed to elicit latent expert knowledge from off-the-shelf models without parameter updates. Second, we introduce a *synthetic-data-driven MLLM adaptation*, where we fine-tune an MLLM to a downstream task using on-the-fly generated synthetic artifacts. The latter approach aims to learn generalized representations of morphing attacks without relying on labeled datasets of specific morphing algorithms, thereby addressing the critical issue of overfitting in current MAD approaches.

3.1. Zero-Shot Forensic Prompting Strategy

Standard prompting strategies (e.g., asking ”Is this face morphed?”) fail to produce reliable results in forensic contexts, often leading to model hallucinations or refusals due to safety alignment [24]. To overcome this issue, we developed a structured prompting methodology grounded in Chain-of-Thought (CoT) reasoning [37], transforming the MLLM from a passive classifier into an active forensic analyst. We additionally condition the MLLM with a forensic analyst system prompt to reduce generic safety refusals and anchor the model in the MAD context.

Structured Analytical Protocol. Our approach moves beyond binary classification by implementing a six-step analytical protocol inspired by NISTIR 8584 [26] guidelines. This protocol explicitly guides the model's attention to anatomical face regions where morphing artifacts typically appear. These six steps are presented as numbered sub-questions in the prompt, and the model must provide a brief textual assessment for each before issuing a final decision. Our proposed prompt sequentially evaluates:

1. *High-Frequency Features:* Scrutinizing fine-grained details around the eyes and lips for ghosting, double edges, or unnatural sharpness discontinuities.
2. *Facial Geometry:* Detecting subtle asymmetries, spatial misalignments, or warping inconsistencies introduced by landmark-based blending operations.
3. *Skin Texture Analysis:* Identifying unnatural smoothing, loss of skin porosity, or texture homogenization commonly observed in attacks generated with deep learning-based methods or heavily retouched imagery.
4. *Boundary Consistency:* Checking for blending artifacts, color mismatch, or edge disruptions at common areas of interest such as hairline, jawline, face contour.
5. *Lighting Coherence:* Verifying consistent illumination direction, shadow placement, and reflectance properties across different facial regions in the image.
6. *Identity Integrity:* Performing a holistic assessment of overall identity coherence, ensuring that facial attributes remain semantically consistent and plausible.

Semantic Scoring and Output Constraints. A key challenge in zero-shot evaluation with MLLMs is the reliable extraction of calibrated, continuous confidence estimates for quantitative performance analysis. In preliminary experiments, we observed that coarse confidence scales (e.g., 0–100, where lower values indicate bona fide images and higher values indicate morphs) induce pronounced score quantization, with predictions collapsing onto a small set of discrete values. This behavior reduces score resolution and degrades the reliability of threshold-based evaluation metrics used to quantify detection error.

To mitigate this issue, we adopt a high-resolution confidence scale ranging from 0 to 10,000, coupled with an explicit semantic interpretation of score intervals. This choice is not intended to increase numerical precision in a statistical sense, but to counteract the tendency of MLLMs to collapse predictions onto a small set of preferred numeric tokens when prompted with coarse ordinal scales. Low-resolution ranges (e.g., 0–100) encourage categorical reasoning and rounded outputs, whereas a larger numeric range supports finer-grained ordinal differentiation. To further stabilize score usage, semantic anchors are defined within the prompt. Scores above 9,000 thus indicate high certainty of a morph, while scores in the 1,000–3,000 range denote likely bona fide images. This guides the model to utilize a full dynamic range and yields smoother score distributions for threshold-based evaluation.

To support automated evaluation and reproducibility, we constrain the model output to a strict JSON schema of the form {`"step1_reasoning": "..."`, `"step1_score": "...", ...`}. For each of the six forensic analysis steps introduced above, the model is required to produce textual reasoning and a score denoting whether the input image represents an attack. The final decision score is obtained by averaging the six step-wise confidence scores and is used for quantitative evaluation and threshold-based decision making. This structured output ensures machine parsability while enforcing a clear separation between reasoning and decision making, thereby improving interpretability and consistency across inference runs. The exact prompts used in our experiments are provided in the Appendix.

3.2. Synthetic-Data-Driven MLLM Adaptation

Zero-shot MLLMs rely solely on broad pretraining and prompt-based reasoning, without task-specific calibration to the subtle visual artifacts characteristic of morphing attacks. Consequently, their sensitivity to fine-grained, low-level inconsistencies, such as localized geometric distortions or frequency-domain artifacts, may be insufficient for reliable morph detection. These limitations motivate targeted adaptation of MLLMs to improve detection accuracy while preserving generalization. We adapt MLLMs using a binary training objective on synthetic data.

Generation of Synthetic Training Data. We adopt a training strategy that simulates typical morphing artifacts rather than using real morphs, similar to [21]. By defining the "attack" class through synthetic perturbations, we force the model to learn generic indicators of manipulation rather than the specific characteristics of actual morphing techniques (e.g., StyleGAN fingerprints). The pipeline for simulation of training data generates training pairs of bona fide and morphed images $(I_{\text{BF}}, I_{\text{M}})$ by processing bona fide inputs I through three separate stages:

- *Pixel-Space Artifact Simulation*: introduces artifacts that simulate irregularities created by landmark-based morphing techniques. Specifically, given an input bona fide image I, this stage first applies a set of randomly parametrized geometrical image transformations ζ:

$$I_{PA} = \zeta(I), \tag{1}$$

  where ζ is randomly sampled from {`Translation`, `ElasticTransform`, `Scaling`}. The pixel-augmented image I_{PA} is blended with the source I using a binary blending mask M:

$$I'_{PA} = I_{PA} \odot a \cdot M + I \odot (1-a) \cdot M, \tag{2}$$

  where a is the blending factor, uniformly sampled from a predefined set $\{0.5, 0.5, 0.5, 0.375, 0.25, 0.125\}$.
- *Frequency-Space Artifact Simulation*: injects structured noise patterns into the Fourier spectrum of the blended image I'_{PA} to mimic the spectral inconsistencies introduced by deep learning morphing techniques, i.e., GANs and diffusion models. Specifically, this stage first creates a random structured geometrical pattern Φ, uniformly chosen to represent one of the following: a symmetrical grid, an asymmetrical grid, a square checkerboard, a circular checkerboard, randomly distributed squares, a set of random lines, or a set of random stripes. The magnitudes of its Fourier transform $F_\Phi = |\text{FFT}(\Phi)|$ are then superimposed on the magnitudes of the Fourier transform of I'_{PA}, $F_{\text{PA}} = |\text{FFT}(I'_{\text{PA}})|$, and transformed back to the image space, by applying the inverse Fourier Transformation FFT^{-1}:

$$I_{\text{FA}} = \text{FFT}^{-1}\big((1-k)F_{\text{PA}} \oplus kF_\Phi\big), \tag{3}$$

  where k is a constant that defines the contribution of Fourier spectra F_{PA} and F_Φ to the summation.
- *Visual Variability Simulation*: focuses on transforming the visual appearance of images to simulate subtle, global visual variations commonly encountered in real-world imagery. Specifically, given an input bona fide image I and a synthetic morph I_{FA}, this stage applies a set of randomly parametrized transformations ψ, to generate a bona fide image I_{BF} and morph I_M:

$$I_{BF} = \psi(I), \qquad I_M = \psi(I_{FA}) \tag{4}$$

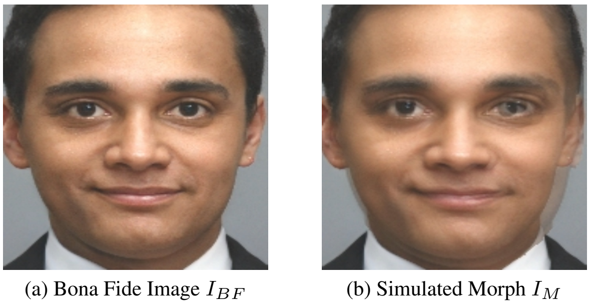
  where ψ is uniformly sampled from {`RGBShift`,`HueSaturationValue`, `RandomBrightnessContrast`, `RandomDownScale`, `Sharpen`} - a set comprising five basic (global) image transformations.

An example of a train pair (I_{BF}, I_M) is shown in Figure 2.

Parameter-Efficient MLLM Adaptation. Full fine-tuning of multi-billion-parameter models is computationally prohibitive and may lead to catastrophic forgetting of pre-trained knowledge. Therefore, we employ LoRA [18] to adapt our MLLM by injecting a small number of trainable parameters while keeping the pre-trained weights frozen. Specifically, for a frozen pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, the weight of the adapted model is

$$W = W_0 + \Delta W, \; \Delta W = BA \tag{5}$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are trainable low-rank matrices with rank $r \ll min(d, k)$.



(a) Bona Fide Image I_{BF} (b) Simulated Morph I_M

Figure 2. Example of a synthetic training image pair used for the MLLM adaptation. (a) represents a bona fide image, (b) is generated via pixel-space and frequency-space artifact simulation.

Importantly, during the adaptation of our MLLM, we apply LoRA adapters to query (q) and value (v) projections of the self-attention layers in *both* the Vision Encoder and the Language Decoder towers. This dual-tower strategy is essential for MAD, as adapting the vision tower allows the model to extract forensic visual cues (e.g., noise patterns) that are likely suppressed in standard pre-training, while adapting the language tower aligns the reasoning engine to the description of typical visual artifacts. Additionally, we append the final aggregated token output of the decoder with a lightweight MLP classification head, optimized using Binary Cross-Entropy (BCE) loss:

$$L_{\text{BCE}} = -\left[y \cdot \log(p) + (1-y) \cdot \log(1-p)\right] \quad (6)$$

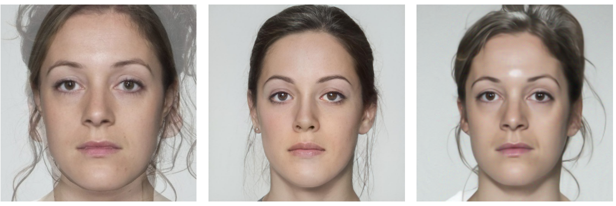
where p is the model's probability to classify an image as a bona fide or a morph, while y is the ground truth label.

The classification head is trained jointly with the LoRA adapters in the vision and language towers, while original MLLM parameters are not updated as they are frozen.

4. Experiments

Experimental MLLMs. For our experiments, we select four different widely used open-source MLLMs: Gemma-3 (27B and 12B) [36], optimized for strong multimodal reasoning with efficient instruction tuning; Qwen2.5-VL 32B [2], known for robust visual understanding and multilingual reasoning; Llama-4-Scout 17B, designed for efficient deployment and fast multimodal inference; and Mistral Small 3.1 24B - a compact yet powerful MLLM emphasizing efficiency and strong language modeling performance. With this selection, we aim to cover various architectural innovations, including Mixture-of-Experts and varying parameter scales, ensuring our findings are robust across different model backbones. In our experiments all models are evaluated in zero-shot settings, while the parameter-efficient LoRA adaption is performed only with Gemma-3 12B, as the best-performing MLLM.

Testing Datasets. To ensure rigorous assessment of MLLMs' MAD performance, we evaluate models across different testing datasets spanning classical landmark-based morphs, GAN-, and diffusion-based attacks. Figure 3 illustrates how visual characteristics of morphs differ depending on the type of the underlying morphing technique. In our experiments, we utilize seven widely used benchmark datasets: FRLL-Morphs, FRGC-Morphs, FERET-Morphs [30], containing morphs generated with morphing algorithms AMSL, FaceMorpher (FM), OpenCV (OCV), StyleGAN2 (SG), and WebMorph (WM), and higher-quality sets MorGAN [8], MIPGAN-II [38], MorDIFF [11], and Greedy-DiM [3]. In addition to these datasets, we also use LMA-DRD [9], included to evaluate performance on printed and scanned images, to assess robustness against re-digitalization noise.



(a) OpenCV (b) StyleGAN2 (c) Greedy-DiM

Figure 3. Examples of a landmark-based (a), a GAN-based (b), and a diffusion-based (c) morph, all three generated using bona fide images from FRLL. The visual characteristics of morphs vary substantially depending on the morphing technique.

Training Data. During the synthetic-data-driven MLLM adaptation, we utilize the bona fide training subset of SMDD [10]. This subset consists of 25,000 synthetic images generated with StyleGAN2 utilized to generate simulated training morphs as described in Section 3.

Evaluation Metrics. In our evaluations, we follow the ISO/IEC 20059:2025 standard[1] by computing the Equal Error Rate (EER), where the Morphing Attack Classification Error Rate (MACER) equals the Bona Fide Sample Classification Error Rate (BSCER). MACER corresponds to the proportion of morphs incorrectly accepted as bona fide, whereas BSCER measures bona fide images falsely rejected as attacks. Beyond EER, in some experiments we also report BSCER at fixed MACER operating points of 1% and 5%. These stricter operating points more accurately reflect real-world identity verification tasks, where low attack-acceptance is essential for system security.

Implementation Details. Our experimental protocol explicitly differentiates between zero-shot evaluation of MLLMs and their adaptation to the MAD task.

During the *zero-shot evaluations*, all models were configured with a temperature of 0.1 to balance output determinism with the creative reasoning required for forensic analysis. Inference was performed using the vLLM engine on a cluster of four NVIDIA RTX 4090 GPUs. To fit memory constraints, Llama-4-Scout utilized 4-bit quantization, while other models used bfloat16. Images were preprocessed following the requirements of each MLLM.

During the lightweight adaptation of Gemma-3 12B,

[1] International Organization for Standardization (ISO). ISO/IEC 20059:2025 — Information technology — Biometric presentation attack detection — Testing and reporting. (This standard supersedes ISO/IEC 30107-3:2017.)

Table 1. Zero-shot MLLM evaluations vs. LoRA-adapted Gemma-3 performance in terms of EER(%). The LoRA-adapted Gemma-3 significantly outperforms all zero-shot baselines, including the classification head trained on top of the pretrained Gemma-3.

Dataset	**Morph type**	**Zero-shot evaluations**				**Gemma-3 + classification head**	
		Mistral Small 3.1	**Llama-4-Scout**	**Qwen2.5-VL**	**Gemma-3**	**Pretrained backbone**	**LoRA adapted**
FRGC-M	FM	49.20	—	42.41	32.19	24.48	**4.98**
	OCV	52.82	—	42.62	43.55	29.72	**9.23**
	SG	50.45	—	59.81	57.06	40.82	**17.32**
FERET-M	FM	53.07	—	34.52	18.20	13.80	**9.30**
	OCV	53.26	—	32.46	19.62	11.73	**7.40**
	SG	48.57	—	34.27	40.94	**27.40**	34.97
FRLL-M	AMSL	42.13	49.63	44.13	25.10	48.62	**12.74**
	FM	37.29	41.50	43.01	13.08	18.15	**0.49**
	OCV	40.24	37.63	39.51	13.33	6.38	**1.47**
	SG	52.06	47.59	26.86	27.39	16.72	**6.83**
	WM	40.78	39.22	41.06	12.88	21.47	**2.37**
LMA-DRD	D	49.01	—	45.40	47.05	31.56	**18.90**
	PS	51.50	—	47.43	43.88	39.74	**28.28**
MorGAN	GAN	55.60	—	48.63	52.58	46.69	**45.26**
	LMA	46.88	—	51.67	52.87	50.00	**23.06**
MIPGAN II	SG	50.24	46.58	20.75	35.56	31.67	**17.92**
Greedy-DiM	DiffAE	50.49	49.93	24.55	**6.15**	11.24	11.78
MorDIFF	DiffAE	47.77	—	45.91	36.13	24.88	**17.33**
Average		48.41	44.58	40.28	32.09	27.50	**14.98**

we leverage LoRA adapters with parameters $r = 16$, and $\alpha = 32$, injected into the query and value projections of all self-attention layers in the vision and the language tower of the MLLM. Weights were optimized for 30 epochs with an effective batch size of 32. To promote stable optimization, we employ a differential learning rate strategy across model components. Specifically, we use a higher learning rate for the randomly initialized classification head (1×10^{-4}), a moderate learning rate for the vision tower (6×10^{-6}), and a substantially lower learning rate for the language tower (3×10^{-7}). This design reflects the differing levels of sensitivity to parameter updates: the classification head requires rapid convergence from scratch, while the vision and language towers—adapted via LoRA—benefit from more conservative updates to preserve pre-trained representations and prevent destabilization of linguistic reasoning. The optimization was performed on two NVIDIA A100 (80GB) GPUs.

Comparison With Existing MADs. We assess the MAD performance of evaluated MLLMs against various established MAD methods. Among supervised baselines, we consider MixFaceNet-MAD [4], Inception-MAD [29], and PW-MAD [9]. As the performance of the supervised methods and their generalization strongly depend on the training data, we train each method on three different datasets, i.e., SMDD, MorGAN, and LMA-DRD, following a protocol established in [14]. In addition to supervised MADs, we also include comparison with unsupervised MADs FIQA-MagFace [16], CNNIQA [16], SPL-MAD [14], and MAD-DDPM [20], conceptually similar self-supervised models SBI [34] and SelfMAD [21], and the foundation model-based method MADation [7].

5. Results

MLLM Evaluation in Zero-Shot Settings. Results obtained during the zero-shot evaluation of selected MLLMs are summarized in Table 1. Among the four selected MLLMs, Gemma-3 27B achieved the best average EER of 32.09%, outperforming the runner-up Qwen2.5-VL by 8.19%. Both Llama-4-Scout and Mistral Small 3.1 performed substantially worse, with an overall EER of 44.58% and 48.4%, respectively. These results demonstrate that MLLMs possess different zero-shot capabilities for detecting morphed faces. The sensitivity of the models to specific morphing techniques also varies considerably. However, MLLMs in general achieve lower attack detection error when tested on artifact-rich morphs, as opposed to the accuracy measured on higher-quality morphed images. Gemma-3, for example, relatively accurately detects blending artifacts in FRLL FaceMorpher, OpenCV, and WebMorph attacks, with an EER ranging between 12.88% and 13.33%. Nevertheless, detection errors are significantly higher on FRGC-StyleGAN morphs (57.1% EER), probably due to the seamless latent-space interpolation performed by the morphing technique StyleGAN, which produces very few perceptible artifacts. Qualitative evaluation examples with corresponding confidence scores and reasoning are shown in Figure 4.

Evaluation of the Adapted MLLM. To isolate the impact of LoRA adaptation, we evaluate Gemma-3 12B using a lightweight, probability-based classifier rather than prompt-derived numeric scores. Specifically, we first attach and train an MLP classification head on top of the frozen (unadapted) MLLM and use its sigmoid output as the morph probability. This bypasses the need to interpret free-form numeric confidence values generated by the language decoder, which are subject to token-level biases and scale-dependent discretization. We then evaluate the LoRA-adapted Gemma-3 12B in the same manner. Results are reported in Table 1. As can be seen, the unadapted MLLM with an added classification head achieves an average EER of 27.50%, outperforming zero-

Table 2. Comparison of the LoRA adapted Gemma-3 with supervised MAD models trained on different datasets in terms of EER(%). Gemma-3 outperforms competitors in terms of average detection accuracy, achieving stable performance across different morph types.

Dataset	**Morph type**	**MixFaceNet-MAD [4]**					**PW-MAD [9]**					**Inception-MAD [29]**					**Gemma-3 [LoRA]**
		D	**PS**	**LMA**	**GAN**	**SMDD**	**D**	**PS**	**LMA**	**GAN**	**SMDD**	**D**	**PS**	**LMA**	**GAN**	**SMDD**	
FRGC-M	OCV	23.81	25.04	31.62	21.11	20.67	57.06	48.60	29.74	53.55	26.45	34.32	13.65	36.17	59.66	19.63	**9.23**
	FM	22.83	23.54	29.38	19.98	18.10	56.00	50.70	30.49	51.61	23.40	34.96	19.71	35.10	56.91	16.06	**4.98**
	SG	32.71	28.68	21.70	21.95	**11.62**	37.38	38.42	16.43	26.62	14.32	41.14	25.85	36.19	47.03	15.26	17.32
FERET-M	OCV	28.12	32.19	31.57	33.86	31.74	37.27	45.29	34.27	43.11	39.93	**6.39**	7.23	42.12	13.62	59.32	7.40
	FM	22.57	29.48	27.90	31.81	23.69	35.16	44.30	28.24	40.40	29.41	**5.17**	6.91	36.53	18.36	46.94	9.30
	SG	29.57	29.02	35.46	39.41	39.85	44.25	45.30	29.70	42.47	47.20	9.03	**7.12**	35.29	15.09	60.05	34.97
FRLL-M	OCV	8.82	13.22	8.91	17.66	4.39	17.33	15.69	13.96	45.59	2.42	13.72	10.76	6.86	55.89	5.38	**1.47**
	FM	7.80	10.97	7.34	15.65	3.87	13.88	15.14	10.92	44.57	2.20	16.62	15.81	6.32	66.14	3.17	**0.49**
	SG	20.07	15.29	13.41	23.51	8.89	29.97	27.64	18.11	48.53	16.64	37.24	19.58	20.56	55.03	11.37	**6.83**
	WM	25.97	29.04	20.61	30.39	12.35	33.78	28.51	35.75	52.43	16.65	57.38	58.32	30.88	77.42	9.86	**2.37**
	AMSL	24.53	27.59	19.24	30.03	15.18	36.25	32.95	34.38	48.52	15.18	49.02	61.44	**9.80**	86.49	10.79	12.74
LMA-DRD	D	15.68	18.03	17.06	25.01	19.42	20.80	25.10	22.34	40.21	17.06	7.64	17.06	15.68	50.77	**15.11**	18.90
	PS	21.77	18.44	27.05	27.05	23.72	26.48	23.72	29.41	44.11	20.39	**11.37**	12.75	22.34	38.42	19.01	28.28
MorGAN	LMA	39.42	**22.89**	10.61	46.42	30.12	34.20	34.14	9.71	34.37	27.31	38.55	31.73	8.43	40.16	28.51	23.06
	GAN	53.01	50.44	42.57	24.90	42.64	52.04	46.59	42.80	8.84	43.78	50.84	38.79	**27.41**	0.40	44.34	45.26
Greedy-DiM	DiffAE	45.10	41.67	40.69	48.04	39.71	17.16	33.82	17.16	15.20	42.16	31.86	51.96	25.98	29.90	56.86	**11.78**
MorDIFF	DiffAE	21.30	23.70	28.83	30.19	20.40	3.21	**0.98**	11.60	16.00	13.80	21.08	21.78	19.41	56.09	15.23	17.33
Average		26.71	26.30	25.21	28.88	21.55	33.21	33.32	25.33	40.46	23.43	28.67	25.48	25.42	47.94	25.70	**14.81**

* Train data: **D** (LMA-DRD - digital), **PS** (LMA-DRD - print&screen), **LMA** (MorGAN - landmark-based), **GAN** (MorGAN - GAN-based), **SMDD**

Table 3. Comparison of the LoRA adapted Gemma-3 with unsupervised MAD models in terms of EER(%) and BSCER at MACER 5% and 10%. Gemma-3 shows competitive average EER, while highlighting complementary strengths with state-of-the-art MADs.

Dataset	**Morph type**	**FIQA-MagFace [16]**			**CNNIQA [16]**			**SPL-MAD [14]**			**MAD-DDPM [20]**			**SBI [34]**			**SelfMAD [21]**			**MADation ViT-B[7]**			**MADation ViT-L[7]**			**Gemma-3 [LoRA]**		
		EER	**5%**	**10%**	**EER**	**5%**	**10%**	**EER**	**5%**	**10%**	**EER**	**5%**	**10%**	**EER**	**5%**	**10%**	**EER**	**5%**	**10%**	**EER**	**5%**	**10%**	**EER**	**5%**	**10%**	**EER**	**5%**	**10%**
FRGC-M	FM	33.82	73.79	62.84	42.84	75.94	66.86	16.91	25.39	21.47	25.62	95.12	90.15	16.68	38.07	26.14	5.59	6.43	2.80	–	–	–	–	–	–	**4.98**	**4.91**	**2.15**
	OCV	33.30	74.71	62.52	43.15	74.64	66.35	20.75	32.50	25.42	28.22	95.12	90.15	15.32	36.31	25.10	**2.59**	**1.14**	**0.41**	–	–	–	–	–	–	9.23	14.11	8.31
	SG	**14.21**	26.46	17.60	36.51	70.34	57.93	16.80	**26.13**	**21.09**	9.02	95.12	90.15	52.90	97.10	94.40	15.84	45.23	25.52	–	–	–	–	–	–	17.32	34.35	25.21
FERET-M	FM	25.14	61.22	44.44	13.23	35.17	19.32	20.42	40.85	27.09	27.98	95.27	90.17	26.47	60.87	52.36	**3.19**	**1.70**	**0.38**	–	–	–	–	–	–	9.30	19.50	8.04
	OCV	26.14	61.50	43.95	20.45	58.60	37.23	25.71	57.45	45.60	31.38	95.27	90.17	28.73	70.08	60.61	**1.13**	**0.57**	**0.38**	–	–	–	–	–	–	7.40	12.75	4.97
	SG	**12.67**	**24.63**	**15.71**	33.84	79.55	66.17	25.33	62.06	49.72	32.14	95.27	90.17	41.83	90.55	82.42	18.14	46.12	32.33	–	–	–	–	–	–	34.97	71.23	61.56
FRLL-M	AMSL	30.94	77.94	66.18	21.61	60.29	39.22	3.26	0.50	0.50	27.13	94.94	90.02	11.76	24.23	16.78	**0.99**	**0.05**	**0.05**	**3.85**	–	**2.89**	7.26	–	10.63	12.74	20.71	14.22
	FM	27.99	73.04	57.35	19.97	57.84	36.76	1.03	0.99	0.99	10.40	95.19	90.38	13.73	36.99	26.10	**0.00**	0.26	0.17	1.35	–	**0.00**	0.74	–	0.98	**0.49**	**0.00**	**0.00**
	OCV	24.73	66.18	53.43	7.53	11.76	4.41	1.88	0.50	0.50	13.76	95.17	90.01	12.25	27.85	18.84	**0.00**	**0.00**	**0.00**	2.97	–	0.49	**0.99**	–	**0.00**	1.47	0.49	**0.00**
	SG	7.53	8.82	5.39	35.92	75.49	68.14	14.65	32.18	24.75	14.32	95.17	90.18	44.61	94.68	90.92	10.34	24.22	12.52	17.21	–	26.69	24.96	–	49.03	**6.83**	**10.29**	**5.39**
	WM	27.19	68.14	55.39	21.54	46.57	33.33	6.39	11.39	3.47	30.30	95.09	90.34	39.22	89.93	83.37	3.45	1.64	0.41	3.42	–	**0.49**	4.07	–	1.47	**2.37**	**1.47**	**0.49**
MIPGAN II	SG	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	22.21	–	47.55	**9.06**	–	**5.39**	17.92	–	23.57
Greedy-DiM	DiffAE	47.00	94.61	85.78	49.40	96.08	93.14	37.72	80.69	71.78	36.10	95.20	89.70	33.82	90.60	81.60	**7.60**	37.60	27.80	–	–	–	–	–	–	11.78	**13.73**	**11.76**
MorDIFF	DiffAE	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**1.10**	–	**0.00**	20.40	–	37.25	17.33	–	26.04
Average (∗)		25.89	59.25	47.55	28.83	61.86	49.07	15.90	30.89	24.36	23.86	95.16	90.13	28.11	63.11	54.89	**5.74**	**13.75**	**8.56**	–	–	–	–	–	–	9.91	16.96	11.84
Average (†)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**7.44**	–	11.16	9.64	–	14.96	8.45	–	**9.96**

* Test data: FRGC-M, FERET-M, FRLL-M, Greedy-DiM; † Test data: FRLL-M, MIPGAN II, MorDIFF

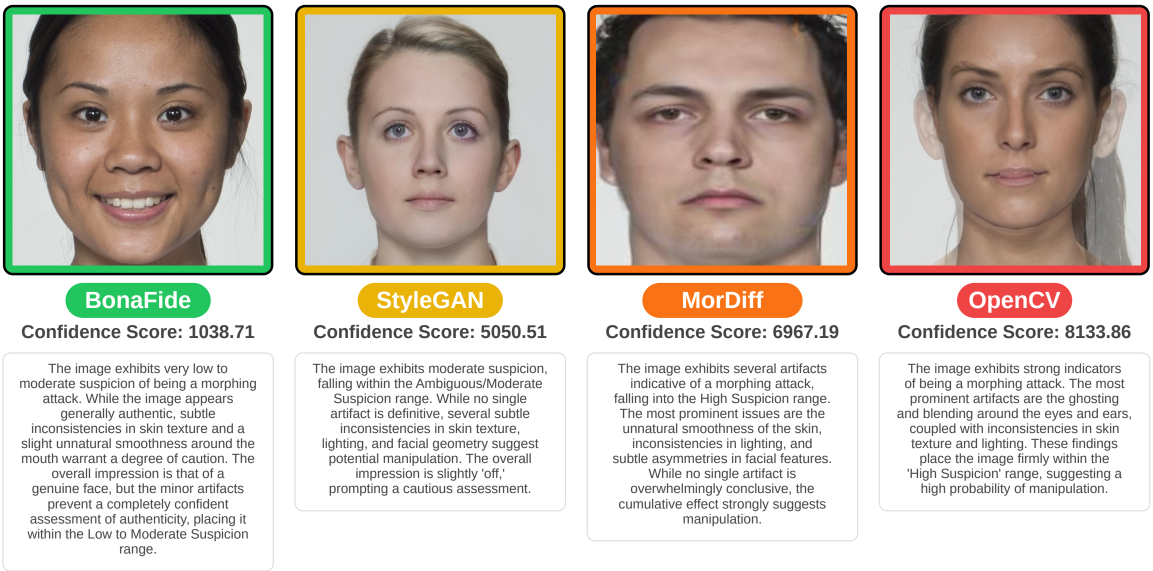
shot evaluations by 4.59 percentage points. However, the LoRA-adapted model significantly reduces the EER to 14.81% (a 46.2% relative improvement), suggesting that unadapted features are not as informative for MAD.

In our experiments, the adaptation was especially beneficial for landmark-based morphs. On FRLL-FaceMorpher, for example, the adapted Gemma-3 12B achieved an EER of 0.49%, a substantial gain over the previously reported zero-shot EER of 13.08%. Similar gains appear on FRLL-OpenCV and FRLL-WebMorph. After the adaptation, the detection accuracy was also significantly improved for some GAN-based morphs, such as FRLL-StyleGAN2 (improved from 27.39% EER to 6.83%). However, GAN-based MorGAN attacks remained challenging even after the MLLM adaptation. To better assess such failures, a further analysis on the impact of the training data used for the adaptation is needed.

Comparison Against Supervised MADs. The empirical comparison of the adapted Gemma-3 12B with existing supervised MADs is given in Table 2. As can be seen, our adapted MLLM outperforms all competitive models, achieving an average EER of 14.81%, a 6.74% improvement over the best supervised MAD baseline MixFaceNet-MAD, trained on SMDD. In addition, we note that supervised baselines exhibit severe overfitting. Inception-MAD trained on GAN morphs, for example, degrades to 55.03% on FRLL-StyleGAN2. In contrast, our Gemma-3 12B-MAD maintains consistent performance across attack types, showing better robustness and generalization across different test data.

Comparison Against Unsupervised MADs. The empirical comparison of the adapted Gemma-3 12B with existing unsupervised and self-supervised MADs is given in Table 3. The HRNet-W18-based SelfMAD achieves the best average EER of 5.74%, outperforming Gemma-3 12B-MAD by 4.17 percentage points. Unlike Gemma-3 12B-MAD, whose vision encoder represents a transformer, SelfMAD extracts visual features with a high-resolution CNN, which is especially good at detecting localized artifacts, important for the MAD task. We also note that while SelfMAD dominates landmark-based attacks, Gemma-3 12B-MAD performs better on certain generative attacks. On Greedy-DiM, Gemma-3 12B-MAD achieves 13.73% BPCER at 5% APCER versus 37.60% for SelfMAD. Moreover, Gemma-3 12B-MAD outperforms both SPL-MAD and MAD-DDPM, showing the strength and the underexplored potential of MLLMs in the context of MAD.

Comparison With Foundation Model-based MADs. We compare against MADation and proprietary MLLMs.


BonaFide
Confidence Score: 1038.71
StyleGAN
Confidence Score: 5050.51
MorDiff
Confidence Score: 6967.19
OpenCV
Confidence Score: 8133.86

The image exhibits very low to moderate suspicion of being a morphing attack. While the image appears generally authentic, subtle inconsistencies in skin texture and a slight unnatural smoothness around the mouth warrant a degree of caution. The overall impression is that of a genuine face, but the minor artifacts prevent a completely confident assessment of authenticity, placing it within the Low to Moderate Suspicion range.

The image exhibits moderate suspicion, falling within the Ambiguous/Moderate Suspicion range. While no single artifact is definitive, several subtle inconsistencies in skin texture, lighting, and facial geometry suggest potential manipulation. The overall impression is slightly 'off,' prompting a cautious assessment.

The image exhibits several artifacts indicative of a morphing attack, falling into the High Suspicion range. The most prominent issues are the unnatural smoothness of the skin, inconsistencies in lighting, and subtle asymmetries in facial features. While no single artifact is overwhelmingly conclusive, the cumulative effect strongly suggests manipulation.

The image exhibits strong indicators of being a morphing attack. The most prominent artifacts are the ghosting and blending around the eyes and ears, coupled with inconsistencies in skin texture and lighting. These findings place the image firmly within the 'High Suspicion' range, suggesting a high probability of manipulation.

Figure 4. Qualitative zero-shot results generated with Gemma-3. Examples illustrate the step-wise forensic reasoning produced by the proposed six-step prompting protocol, together with the final aggregated confidence score for one bona fide and three morphing attacks.

Prior work [39] reports GPT-4 Turbo achieves 37.0% EER on MIPGAN-II in zero-shot evaluation. Our LoRA-tuned Gemma-3 12B achieves 17.92%, demonstrating that domain-specific adaptation outperforms larger proprietary models. Compared to MADation ViT-B (7.44% average EER on FRLL/MIPGAN-II/MorDIFF), Gemma-3 12B performs comparably (8.45%). At strict thresholds, Gemma-3 achieves 9.96% BPCER at 10% APCER versus 14.96% for MADation ViT-L (Table 3).

6. Conclusion

This paper explored the potential of open-source Multimodal Large Language Models (MLLMs) for single-image morphing attack detection (S-MAD) under strict cross-dataset evaluation, addressing both generalization and interpretability, two longstanding challenges in biometric security. We explored MLLMs through two paradigms: structured zero-shot forensic prompting and parameter-efficient, synthetic-data-driven adaptation.

First, we demonstrated that carefully designed multi-step forensic prompting, inspired by established forensic analysis guidelines, can effectively elicit latent morphing-related knowledge from pretrained MLLMs without task-specific training. The proposed Chain-of-Thought protocol significantly improves zero-shot detection performance compared to naive prompts, while simultaneously producing human-readable, region-level semantic explanations. Notably, zero-shot Gemma-3 exhibits competitive performance on diffusion-based morphs, outperforming specialized CNN-based detectors in certain cases, highlighting the complementary forensic sensitivity of MLLMs to emerging face morphing attack types.

Second, we showed that parameter-efficient LoRA adaptation, guided by privacy-preserving synthetic artifacts, substantially enhances MLLM detection accuracy and cross-dataset robustness. The adapted Gemma-3 12B-MAD model achieves a strong average EER across eight diverse benchmarks, outperforming widely used supervised and unsupervised MAD methods and approaching the performance of state-of-the-art self-supervised and foundation model–based detectors. This confirms that MLLMs can be effectively adapted to the MAD task without reliance on real-world datasets or proprietary models.

Despite these advances, important limitations remain. Zero-shot MLLM performance, while inherently interpretable and informative, is not yet sufficient for deployment in high-security operational settings with strict accuracy requirements. Conversely, LoRA-adapted models currently operate as binary classifiers and do not retain the rich, structured forensic explanations available in zero-shot inference. Additionally, performance degradation on low-resolution and re-digitized images highlights the need for improved robustness to adverse data acquisition.

These findings point to a promising future direction: conversational and multi-objective fine-tuning of MLLMs, enabling models to jointly deliver classifier-level accuracy and structured, step-by-step forensic reasoning. Such models could effectively bridge the gap between transparency and performance, positioning MLLM-based MAD systems as trustworthy, explainable, and operationally viable tools for biometric security applications.

Acknowledgements

The research presented in this work was supported in part by the ARIS Research Program P2-0250(B) “Metrology and Biometric Systems”, and the ARIS Research Project J2-50065 “DeepFake DAD”, as well as ARIS Research Program P2-0103 “Knowledge Technologies”, young researcher grant no. PR-13689, grant no. PR-12379, the EU project “LLMs4EU” (Contract No. 101198470) and the “ANOZ” project ON-2309.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint:2303.08774*, 2023. 2, 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint:2502.13923*, 2025. 5

[3] Zander W. Blasingame and Chen Liu. Greedy-DiM: Greedy Algorithms for Unreasonably Effective Face Morphs. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2024. 2, 5

[4] Fadi Boutros, Naser Damer, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. MixFaceNets: Extremely Efficient Face Recognition Networks. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2021. 2, 6, 7

[5] Eduarda Caldeira, Pedro C Neto, Tiago Gonçalves, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. Unveiling the Two-Faced Truth: Disentangling Morphed Identities for Face Morphing Detection. In *31th European Signal Processing Conference (EUSIPCO)*, 2023. 3

[6] Eduarda Caldeira, Fadi Boutros, and Naser Damer. MAD-PromptS: Unlocking Zero-Shot Morphing Attack Detection with Multiple Prompt Aggregation. In *Proceedings of the 1st International Workshop & Challenge on Subtle Visual Computing*, 2025. 1, 3

[7] Eduarda Caldeira, Guray Ozgur, Tahar Chettaoui, Marija Ivanovska, Peter Peer, Fadi Boutros, Vitomir Struc, and Naser Damer. MADation: Face Morphing Attack Detection with Foundation Models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025. 1, 2, 3, 6, 7

[8] Naser Damer, Alexa Moseguí Saladié, Andreas Braun, and Arjan Kuijper. MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018. 2, 5

[9] Naser Damer, Noémie Spiller, Meiling Fang, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. PW-MAD: Pixel-Wise Supervision for Generalized Face Morphing Attack Detection. In *International Symposium on Visual Computing (ISVC)*, 2021. 1, 2, 5, 6, 7

[10] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly Synthetic Data for the Development of Face Morphing Attack Detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 5

[11] Naser Damer, Meiling Fang, Patrick Siebke, Jan Niklas Kolf, Marco Huber, and Fadi Boutros. MorDIFF: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Diffusion Autoencoders. In *11th International Workshop on Biometrics and Forensics (IWBF)*, 2023. 1, 2, 5

[12] Luca Debiasi, Christian Rathgeb, Ulrich Scherhag, Andreas Uhl, and Christoph Busch. PRNU Variance Analysis for Morphed Face Image Detection. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018. 2

[13] Lisa DeBruine and Benedict Jones. Face Research Lab London Set, 2017. 2

[14] Meiling Fang, Fadi Boutros, and Naser Damer. Unsupervised Face Morphing Attack Detection via Self-paced Anomaly Detection. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2022. 1, 2, 3, 6, 7

[15] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. *On the Effects of Image Alterations on Face Recognition Accuracy*. Springer International Publishing, 2016. 1

[16] Biying Fu and Naser Damer. Face morphing attacks and face image quality: The effect of morphing and the unsupervised attack detection by quality. *IET Biometrics*, 11 (5), 2022. 2, 6, 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR)*, 2022. 3, 4

[19] Marco Huber, Fadi Boutros, Anh Thi Luu, Kiran B. Raja, Raghavendra Ramachandra, Naser Damer, Pedro C. Neto, Tiago Gonçalves, Ana F. Sequeira, Jaime S. Cardoso, et al. SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-aware Synthetic Training Data. In *International Joint Conference on Biometrics (IJCB)*, 2022. 1

[20] Marija Ivanovska and Vitomir Štruc. Face Morphing Attack Detection with Denoising Diffusion Probabilistic Models. In *11th International Workshop on Biometrics and Forensics (IWBF)*, 2023. 2, 3, 6, 7

[21] Marija Ivanovska, Leon Todorov, Naser Damer, Deepak Kumar Jain, Peter Peer, and Vitomir Štruc. SelfMAD: Enhancing Generalization and Robustness in Morphing Attack Detection via Self-Supervised Learning. In *IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, 2025. 1, 2, 3, 4, 6, 7

[22] Marija Ivanovska, Leon Todorov, Peter Peer, and Vitomir Štruc. SelfMAD++: Self-supervised foundation model with local feature enhancement for generalized morphing attack detection. *Information Fusion*, 2026. 2, 3

[23] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A Universal Representation for Face Recognition and Quality Assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[24] Kartik Narayan, VS Vibashan, and Vishal M. Patel. FaceXBench: Evaluating Multimodal LLMs on Face Understanding (T-BIOM). *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2026. 3

[25] Pedro C Neto, Tiago Gonçalves, Marco Huber, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. OrthoMAD: Morphing Attack Detection Through Orthogonal Identity Disentanglement. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2022. 3

[26] Mei Ngan and Patrick Grother. Face Analysis Technology Evaluation (FATE) MORPH Part 4B: Considerations for Implementing Morph Detection in Operations. NIST Interagency Report (NISTIR) 8584, National Institute of Standards and Technology, Gaithersburg, MD, 2025. 3

[27] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29 (1), 1996. 2

[28] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5), 1998. 2

[29] Raghavendra Ramachandra, Sushma Venkatesh, Kiran Raja, and Christoph Busch. Detecting Face Morphing Attacks with Collaborative Representation of Steerable Features. In *3rd International Conference on Computer Vision and Image Processing (CVIP)*, 2020. 6, 7

[30] Eklavya Sarkar, Pavel Korshunov, Laurent Colbois, and Sébastien Marcel. Vulnerability Analysis of Face Morphing Attacks from Landmarks and Generative Adversarial Networks. *arXiv preprint:2012.05344*, 2020. 5

[31] Ulrich Scherhag, Luca Debiasi, Christian Rathgeb, Christoph Busch, and Andreas Uhl. Detection of Face Morphing Attacks Based on PRNU Analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 1(4), 2019. 2

[32] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. Face Recognition Systems Under Morphing Attacks: A Survey. *IEEE Access*, 7, 2019. 2

[33] Ria Shekhawat, Hailin Li, Raghavendra Ramachandra, and Sushma Venkatesh. Towards Zero-Shot Differential Morphing Attack Detection with Multimodal Large Language Models. In *IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, 2025. 1

[34] Kaede Shiohara and Toshihiko Yamasaki. Detecting Deepfakes with Self-Blended Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 7

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[36] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 Technical Report. *arXiv preprint:2503.19786*, 2025. 5

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *International Conference on Neural Information Processing Systems (NIPS)*, 35, 2022. 3

[38] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Bylappa Raja, Naser Damer, and Christoph Busch. MIPGAN: Generating strong and high quality morphing attacks using identity prior driven GAN. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 3(3), 2021. 2, 5

[39] Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. ChatGPT Encounters Morphing Attack Detection: Zero-Shot MAD With Multi-Modal Large Language Models and General Vision Models. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 2025. 2, 3, 8